

Transcriptome-Based Deep Learning for Predicting Tumour Aggressiveness

Masoud Karimi

CSI4900 Honours Project

Faculty of Engineering, University of Ottawa

Abstract: Predicting tumor aggressiveness from transcriptomic data is of significant clinical interest, but predictive performance is expected to vary across cancer types. Using RNA-seq profiles and matched histological grade labels from *The Cancer Genome Atlas (TCGA)*, we defined a binary classification task (G3–G4 vs G1–G2) across 3,948 samples and 19,313 gene-expression features from 11 cancers. Random Forest, XGBoost, and multilayer perceptron models were evaluated using cancer-specific cross-validation, with feature selection performed within training folds to prevent data leakage, and validation-tuned thresholds. Due to class imbalance, balanced accuracy was the primary metric, supported by area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. Results show that cancer-specific modeling is more reliable than a single pan-cancer approach, with performance varying by cancer type and class balance. Uterine carcinoma was the strongest case, while Head and Neck cancer provided a more challenging contrast. Gene-level analysis identified grade-associated candidates (e.g., AURKA, IHH), which warrant further validation as biomarkers rather than constituting causal claims.

Keywords — Tumor grade classification, RNA-seq, TCGA, cancer-specific modeling, imbalanced classification, feature selection, machine learning

I. INTRODUCTION

Tumor grade is a clinically meaningful indicator of aggressiveness because it summarizes how abnormal and poorly differentiated a tumor appears under pathological review. In practice, however, tumors with similar histological labels can still differ in their underlying molecular programs, which raises the question of whether transcriptomic measurements can recover signals associated with more aggressive disease states. RNA sequencing is especially attractive for this task because it captures genome-wide expression patterns that may reflect proliferation, differentiation, stress response, and pathway activation more directly than single-gene biomarkers.

This question is especially relevant in cancers where grade influences prognosis, treatment intensity, or follow-up decisions. At the same time, it is biologically risky to assume that a single molecular classifier should behave uniformly across all tumor types. Cancer cohorts differ in sample size, class balance, baseline heterogeneity, and the biological meaning of grade itself. That heterogeneity became one of the central findings of this project: what initially looked like a general deep-learning classification problem became, after more careful evaluation, a cancer-specific model-selection problem.

The project therefore asks two linked questions. First, can RNA-seq expression profiles distinguish lower-grade from higher-grade tumors within individual cancer cohorts? Second, if a predictive signal exists, is it shared strongly enough across diseases to justify a single pan-cancer model, or is it more reliable to train and evaluate models separately by cancer type? This report answers those questions by comparing tree-based baselines and neural models under fold-safe feature selection and balanced-accuracy-aware evaluation.

The uterine case is particularly informative because endometrial carcinoma is both clinically important and biologically heterogeneous. TCGA showed that endometrial carcinomas span multiple molecular classes and that a subset of histologically high-grade endometrioid tumors resembles serous-like disease, emphasizing that aggressive uterine tumors need not share a single transcriptomic profile [1]. That background makes uterine cancer a strong test case for whether cancer-specific transcriptomic modeling can identify interpretable grade-associated signals.

II. DATA AND PREPROCESSING

The study used TCGA RNA-seq expression data and matched clinical metadata, which were merged by patient identifier to align gene-expression features with histological grade and primary disease labels. After preprocessing, the final modeling table contained gene-expression features, a binary grade label, and the cancer-type field `X_primary_disease`. The working dataset contained 3,948 samples and 19,313 features after grade cleaning, cohort filtering, and sample alignment.

Histological grade was converted into a binary classification target to focus the analysis on clinically meaningful aggressiveness differences. Grades ‘G3’ and ‘G4’ were grouped as the positive class, representing higher-grade tumors, while grade ‘G1 and G2’ were grouped as the negative class. Ambiguous values such as ‘GX’, ‘Unknown’, and discrepancy labels were removed. Low-grade and high-grade text labels were normalized during preprocessing before the final binary mapping was applied.

The filtering stage was substantial and directly shaped the final dataset. Of the 9,627 original samples, 5,581 were removed due to missing histological-grade labels, 95 were excluded because of invalid grade values, and 2 additional samples could not be mapped to the G1–G4 scale after normalization. Following alignment with the expression data, the final dataset contained 3,948 samples, indicating that approximately 59% of the original cohort was excluded. Importantly, this reduction was driven primarily by the availability of grade labels rather than limitations in RNA-seq data.

TABLE I: RETAINED CANCER COHORTS AFTER FILTERING AND BINARY GRADE ASSIGNMENT

Disease	Total Samples	Low Grade	High Grade
Uterine Carcinoma	521	211	310
Brain Lower Grade	500	242	258

Head & Neck	475	356	119
Kidney Clear Cell	403	178	225
Bladder Urothelial	402	21	381
Stomach Adenocarcinoma	359	143	216
Ovarian Serous	355	42	313
Liver Hepatocellular	354	221	133
Cervical & Endocervical	258	144	114
Pancreatic Adenocarcinoma	160	111	49
Esophageal Carcinoma	125	81	44
Cholangiocarcinoma	36	16	20

The retained cohorts varied substantially in size and class balance. Uterine corpus endometrioid carcinoma included 521 samples with a 310/211 high-to-low grade split. Bladder carcinoma was highly skewed toward high grade, with only 21 lower-grade cases, whereas Head and Neck squamous cell carcinoma was imbalanced in the opposite direction (119 positive, 356 negative). Brain lower grade glioma was comparatively balanced (258/242). Due to low sample size, cholangiocarcinoma was chosen to be omitted from the training models and analysis, making it 11 cancers total.

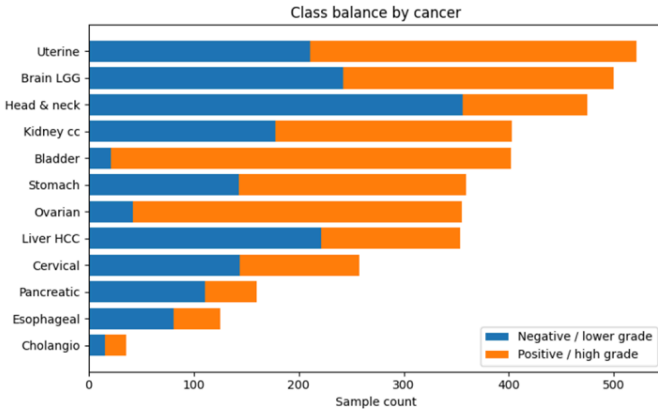


Fig. 1 Class balance of high and low grade by cancer type

III. METHODS

The final methodological principle of the project is cancer-specific model selection. Rather than training one global classifier across all diseases, each retained cancer type was modeled separately so that differences in biology, sample size, and class imbalance were not averaged away. This change in direction came after earlier global experiments produced acceptable overall AUC values but transferred inconsistently across cancers under leave-one-cancer-out style evaluation.

Because the number of expression features was much larger than the number of samples in each cancer cohort, feature selection was essential. The main approach used SelectKBest with ANOVA f_{classif} , and later experiments also considered mutual information as a more flexible non-linear ranking strategy. Feature selection was performed inside each training fold to avoid test leakage. The main MLP sweeps evaluated multiple feature counts, including 250, 500, 1000, 2000, and 5000, which allowed the project to study not only which model performed best, but also how sensitive each cancer was to the size of the retained gene panel.

Three model families formed the basis of the final comparison: tree-based methods and a neural network approach. Random Forest and XGBoost were used as strong baseline models, while a multilayer perceptron (MLP) served as the primary deep learning model. The Random Forest was configured with 600 trees, square-root feature subsampling, and a minimum leaf size of 2. XGBoost used 500 estimators with shallow trees and class weighting to address imbalance.

The MLP was implemented as a regularized feedforward network with two hidden layers (512 and 128 units), using ReLU activations, batch normalization, and 0.35 dropout, followed by a sigmoid output layer. Training employed AdamW optimization, balanced class weighting, early stopping, and validation AUC monitoring.

Autoencoder variants were explored during the project, but they were not retained because their gains were selective rather than consistent.

Evaluation used five-fold cross-validation and emphasized metrics that remain meaningful under imbalance. ROC-AUC was used to measure ranking quality independent of a single threshold, while balanced accuracy was treated as the primary classification metric because it weights sensitivity and specificity equally.

In addition to reporting balanced accuracy at the default threshold, the project also tuned thresholds on validation data within each fold and then applied those thresholds to the held-out fold. This validation-tuned balanced accuracy became the main reporting metric because it better reflects whether a model can classify both high-grade and lower-grade tumors fairly.

The feature-level results support a cautious but meaningful biological interpretation. In uterine cancer, the model consistently highlighted proliferation-associated genes in high-grade tumors and more differentiated, hormone-related signals in lower-grade tumors. In head and neck cancer, the most stable features were consistent with squamous tissue biology. These patterns increase confidence that the models captured real biological structure, but they do not establish mechanisms. Interpretation should therefore be framed as hypothesis generation for follow-up studies rather than as clinical biomarker discovery.

TABLE II: SUMMARY OF MODEL CONFIGURATIONS AND EVALUATION STRATEGY

Component	Configuration
Baseline Model	Random Forest, XGBoost
Neural model	Direct multilayer perceptron
Feature selection	SelectKBest with ANOVA F test, mutual information explored separately

Feature counts tested	250, 500, 1000, 2000, 5000
Cross validation	Five-fold cross-validation
Main metric	Tuned balanced accuracy
Supporting metrics	ROC-AUC, sensitivity, specificity, PR-AUC
Threshold selection	Tuned on validation data within each fold
Interpretation tools	Feature rankings, cross-validation stability, SHAP

IV. RESULTS

Before discussing model performance, it is important to explain why ordinary accuracy was insufficient for this study. Several cohorts exhibited severe class imbalance, meaning that a model could achieve high accuracy by simply predicting the majority class. The bladder cohort illustrates this clearly: it contained 381 high-grade samples and only 21 lower-grade samples (94.8% positive), so a naïve classifier predicting all samples as high grade would already achieve 94.8% accuracy. This demonstrates how standard accuracy can substantially overestimate model performance when minority-class errors are not adequately penalized.

Balanced accuracy was therefore more aligned with the biological and clinical objective of the task. A useful classifier should detect aggressive tumors without completely sacrificing performance on the lower-grade group, especially when the minority class is already underrepresented. This choice of metric also made cross-cancer comparisons fairer because it reduced the extent to which performance was driven by prevalence alone.

Across cancers, no single model family dominated every setting. The tree baselines achieved a mean AUC of 0.741 and a mean tuned balanced accuracy of 0.673, establishing that the problem had real signal even without deep learning. The ANOVA-based MLP was the most broadly competitive deep model, with a mean AUC of 0.749 and a mean tuned balanced accuracy of 0.691. Mutual-information-based MLP results were more cancer-specific, with a mean AUC of 0.742 and a mean tuned balanced accuracy of 0.688, and

were most useful as a targeted improvement rather than a universal replacement.

The best-performing method also changed depending on the cancer and the metric being emphasized. By tuned balanced accuracy, the strongest overall result came from uterine carcinoma, where the selected MLP achieved AUC = 0.922 and tuned balanced accuracy = 0.859. Head and Neck carcinoma also benefited from cancer-specific modeling, reaching AUC = 0.780 and tuned balanced accuracy = 0.727 with an ANOVA-based MLP. In other cancers, such as stomach, tree models remained best, reinforcing the broader conclusion that model choice should be treated as cancer-specific rather than universal.

The comparison between ANOVA and mutual information also supports a nuanced conclusion. Mutual information was clearly useful in selected cohorts such as uterine and esophageal carcinoma, and it improved tuned balanced accuracy in liver HCC. However, ANOVA remained stronger in Head and Neck, ovarian, and pancreatic cancer, and near-tied in several other cohorts. These findings indicate that more flexible feature-ranking methods do not necessarily produce stronger biological signal. Rather, the usefulness of nonlinear feature ranking depends on the underlying structure of the cancer cohort.

Best Tuned Balanced Accuracy by Cancer

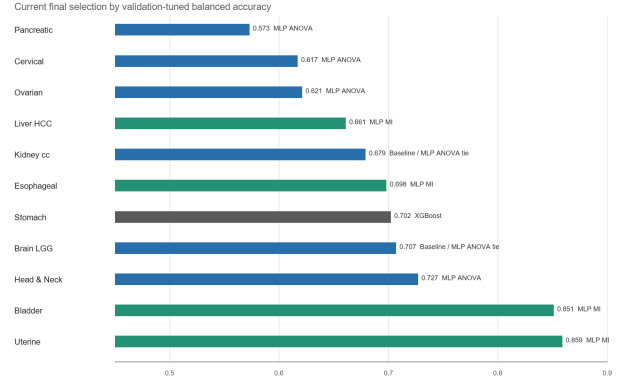


Fig. 2 Best tuned balanced accuracy by cancer and selected model

To interpret the strongest results, the highest-performing cohorts were further analyzed using cross-validation feature stability and SHAP-based importance summaries. Cross-validated metrics were used for formal performance reporting, while SHAP was used to

identify genes most strongly associated with model predictions. These findings are presented as model explanations rather than causal evidence.

V. BIOLOGICAL INTERPRETATION OF HIGH PERFORMING COHORTS

Uterine corpus endometrioid carcinoma is the strongest reliable interpretation case in the project because it combines strong predictive performance, adequate sample size, and a comparatively balanced class split. In the main cross-validation analysis, the uterine cohort reached $AUC = 0.922$ and tuned balanced accuracy = 0.859, making it one of the best-performing cancers in the full comparison.

This result is scientifically meaningful because the model does not appear to be relying on a single unstable artifact. Instead, the uterine analysis consistently points to a proliferation-associated signature in high-grade tumors and a contrasting differentiated uterine program in lower-grade tumors. Genes such as AURKA, FBXO27, C19orf40, and PLAC1 were elevated in the high-grade direction, while genes such as IHH and ADAM28 shifted in the lower-grade direction. The combination of cross-validation stability, SHAP importance, and large expression differences makes this case more convincing than a purely performance-based ranking would.

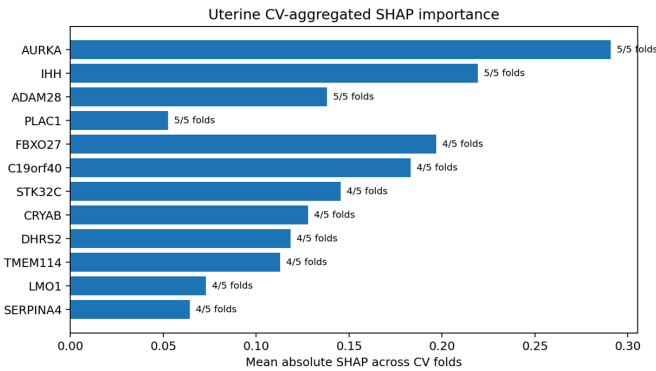


Fig. 3 Feature stability across 5-fold cross-validation based on SHAP importance.

Among the uterine genes, AURKA is one of the most defensible to discuss in the main body. In the project results, AURKA was higher in high-grade tumors and appeared in 5/5 SHAP folds. This aligns with prior endometrial cancer literature reporting that AURKA overexpression is associated with poorer differentiation, higher tumor grade, and

more aggressive behavior [6]. This supports interpreting AURKA as a proliferation-linked marker captured by the model rather than as a stand-alone causal biomarker.

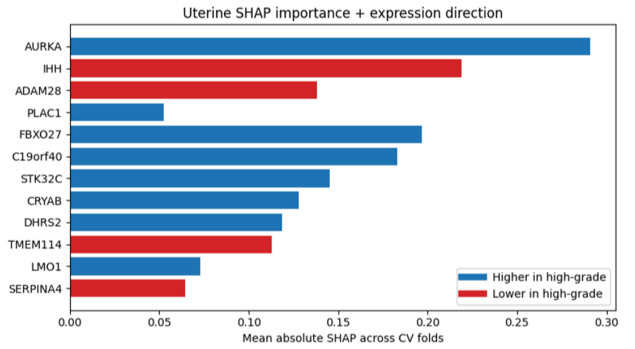


Fig. 4 Expression direction of top stable genes

IHH provides a useful biological counterpoint because it shifts in the opposite direction. In the project analysis, IHH was lower in high-grade uterine tumors and was supported across 5/5 SHAP folds. In uterine biology, Indian hedgehog (IHH) is a progesterone-responsive gene that helps carry out progesterone's effects in the uterus [4], [5]. This suggests that reduced IHH expression in high-grade tumors may reflect loss of progesterone-related signaling, although this remains an indirect interpretation.

Loss of progesterone-receptor expression in human endometrioid carcinoma has been associated with higher histological grade and poorer outcomes [3], while broader reviews describe progesterone signaling as a central tumor-suppressive and differentiation-promoting influence in the endometrium [2]. In this context, reduced IHH expression in high-grade tumors may reflect disruption of a differentiated, hormone-responsive uterine program, although this interpretation remains a model-based biological hypothesis rather than direct evidence.

Head and Neck squamous cell carcinoma is the best contrast case because it is more imbalanced and more difficult than uterine carcinoma, but it still yields coherent features. The main cross-validation result reached $AUC = 0.780$ and tuned balanced accuracy = 0.727. These values are weaker than the uterine result, but they remain strong enough to support a biologically interpretable second case.

The stable features for Head and Neck cancer were enriched for genes associated with squamous differentiation, keratinization, epithelial structure, and inflammatory context. SHAP analysis showed that *SPRR1B*, *SPRR1A*, *DSC1*, *KRT17*, *S100A12*, and *TGM5* were consistently lower in high-grade tumors. Although overall performance was moderate, the model recovered coherent tissue-specific biological signals. This supports the broader conclusion that the transcriptomic interpretation of tumor grade is cancer-specific rather than uniform across disease

Not every high-performing number should be emphasized equally. Bladder carcinoma reached very high AUC and tuned balanced-accuracy values, but the cohort had only 21 lower-grade cases in the full dataset. That makes specificity and threshold stability much more fragile than the headline metric alone suggests.

VI. DISCUSSION

The discussion begins with the central finding that transcriptome-based grade prediction is more reliable when approached as a cancer-specific modeling task rather than a single pan-cancer classification problem. This conclusion emerged from evaluation rather than being imposed a priori. While global performance initially appeared promising, disease-specific analysis showed that differences in sample size, class balance, and underlying biology substantially alter the structure of the problem, limiting the reliability of a universal classifier.

A second important conclusion is that deep learning was competitive but not uniformly dominant. The ANOVA-based MLP was the strongest broad MLP configuration, and deep models performed especially well in cancers such as uterine and Head and Neck. However, tree baselines still won or tied in several cohorts, including stomach and kidney. This is a useful technical result because it argues against treating neural models as automatically superior in moderate-sized omics datasets.

The feature-level results support a cautious but meaningful biological interpretation. In uterine cancer, the model consistently highlighted

proliferation-associated genes in high-grade tumors and more differentiated, hormone-related signals in lower-grade tumors. In head and neck cancer, the most stable features were consistent with squamous tissue biology. These patterns increase confidence that the models captured real biological structure, but they do not establish mechanisms. Interpretation should therefore be framed as hypothesis generation for follow-up studies rather than as clinical biomarker discovery.

VII. LIMITATIONS

The strongest limitation is the lack of external validation. All formal results in this report are TCGA-based, so even the most convincing cancer-specific signals should still be treated as internal evidence rather than deployment-ready biomarkers. A future independent cohort would be necessary to test whether the same features and thresholds generalize across platforms, institutions, and preprocessing pipelines.

A second limitation is that some cohorts remain unstable because of minority-class scarcity. Bladder is the clearest example: despite excellent AUC values, the lower-grade class is too small to support strong specificity claims. More generally, any cohort with a very small minority class can produce threshold-sensitive estimates that vary meaningfully across folds.

A third limitation is interpretability scope. The feature explanations use cross-validation stability, expression direction, and SHAP-style summaries, which are useful for model understanding, but they are still correlational. Gene-level results can suggest plausible biology, especially in the uterine case, yet they do not prove causation or identify therapeutic targets on their own.

VIII. FUTURE WORK

The most important next step is external validation of the final per-cancer selected models on an independent cohort. If a second uterine dataset reproduces the main result, especially the combination of strong balanced accuracy and stable feature direction, the overall claim of cancer-specific transcriptomic grade prediction becomes much stronger.

The second priority is pathway-level interpretation. Rather than relying only on single-gene rankings, future work should examine whether the stable features identified in uterine and Head and Neck cancers group into coherent biological processes. This would make the interpretation more robust and less sensitive to individual gene rankings, and help connect the model outputs to higher-level biological meaning.

Additional technical work could further improve performance and reporting quality, particularly for weaker cancer cohorts. Future work should explore alternative modeling strategies, including more advanced deep-learning and representation-learning approaches tailored to imbalanced or low-sample settings. Autoencoder-based methods were evaluated but did not produce consistent gains, suggesting that more targeted approaches may be needed. Complementary improvements such as calibration analysis, repeated cross-validation, and more stable feature selection would also help distinguish true signal from variability.

IX. CONCLUSION

This project shows that RNA-seq-based tumor-grade prediction is not best understood as a one-model-fits-all deep-learning problem. Instead, the evidence supports a cancer-specific pipeline in which feature selection, threshold tuning, and model comparison are performed separately within each disease cohort. That framing is better aligned with both the biological heterogeneity of cancer and the statistical heterogeneity of the dataset.

Among the evaluated cohorts, uterine corpus endometrioid carcinoma provided the clearest and most reliable signal, combining strong predictive performance with stable, interpretable

feature patterns. Head and Neck squamous cell carcinoma served as a useful contrast case that remained biologically coherent under a more difficult imbalance setting. Together, these cases support the conclusion that transcriptomic models can recover grade-associated molecular structure in selected cancers, but that claims should remain cancer-specific and carefully calibrated to class balance and cohort stability.

X. ACKNOWLEDGEMENT

I thank Dr. Arvind S. Mer and Amir Ebrahimi for supervision and guidance throughout this project.

REFERENCES

- [1] The Cancer Genome Atlas Research Network, "Integrated genomic characterization of endometrial carcinoma," **Nature**, vol. 497, no. 7447, pp. 67-73, 2013.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC3704730/>
- [2] S. Yang, K. W. Thiel, and K. K. Leslie, "Progesterone: the ultimate endometrial tumor suppressor," **Trends in Endocrinology and Metabolism**, vol. 22, no. 4, pp. 145-152, 2011.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC4062362/>
- [3] S. Saito et al., "Progesterone receptor isoforms as a prognostic marker in human endometrial carcinoma," **Cancer Science**, vol. 97, no. 12, pp. 1308-1314, 2006.
<https://pubmed.ncbi.nlm.nih.gov/16999816/>
- [4] K. Lee et al., "Indian hedgehog is a major mediator of progesterone signaling in the mouse uterus," **Nature Genetics**, vol. 38, no. 10, pp. 1204-1209, 2006.
<https://pubmed.ncbi.nlm.nih.gov/16951680/>
- [5] L. Simon et al., "Stromal progesterone receptors mediate induction of Indian Hedgehog (IHH) in uterine epithelium and its downstream targets in uterine stroma," **Endocrinology**, vol. 150, no. 8, pp. 3871-3876, 2009.
<https://pubmed.ncbi.nlm.nih.gov/19372202/>
- [6] K. Umene et al., "Aurora kinase A has a significant role as a therapeutic target and clinical biomarker in endometrial cancer," **International Journal of Oncology**, vol. 46, no. 4, pp. 1498-1506, 2015.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC4356503/>